# Install necessary packages if not already installed

install.packages("dplyr")

install.packages("caret")

install.packages("magrittr")

# Load necessary libraries

library(dplyr)

library(caret)

library(magrittr)

# Load the dataset

credit\_data <- read.csv("/Users/madanshiva/Downloads/CREDIT application\_record.csv")

# Display the structure of the dataset

str(credit\_data)

# Display summary statistics of the dataset

summary(credit\_data)

# Check for missing values

missing\_values <- sapply(credit\_data, function(x) sum(is.na(x)))

print(missing\_values)

# Handle missing values by filling with 'Unknown' for 'OCCUPATION TYPE'

credit\_data$OCCUPATION.TYPE[is.na(credit\_data$OCCUPATION.TYPE)] <- "Unknown"

# Remove duplicates

credit\_data <- credit\_data %>% distinct()

# Encode categorical variables using one-hot encoding

dummies <- dummyVars(~., data = credit\_data)

credit\_data\_encoded <- predict(dummies, newdata = credit\_data)

credit\_data\_encoded <- as.data.frame(credit\_data\_encoded)

# Display the first few rows of the encoded dataset

head(credit\_data\_encoded)

## Question 1: What is the average total household income for each education level?

# Calculate average total household income for each education level

# Ensure dplyr is loaded

library(dplyr)

# Use the correct column name as per the dataset

average\_income\_by\_education <- credit\_data %>%

group\_by(EDUCATION.LEVEL) %>%

summarise(Average\_Income = mean(TOTAL.HOUSEHOLD.INCOME, na.rm = TRUE))

# Format the income in dollars

average\_income\_by\_education$Average\_Income <- paste0("$", formatC(average\_income\_by\_education$Average\_Income, format = "f", big.mark = ",", digits = 2))

print(average\_income\_by\_education)

## Question - 2 What is the distribution of the number of children in the dataset?

install.packages("ggplot2")

# Load ggplot2 library

library(ggplot2)

# Plot the distribution of the number of children

ggplot(credit\_data, aes(x = NO.OF.CHILDREN)) +

geom\_bar(fill = "blue", color = "black") +

labs(title = "Distribution of Number of Children", x = "Number of Children", y = "Count")

##Question -3 What is the average age of applicants based on their housing type?

library(dplyr)

# Calculate the average age of applicants based on their housing type

average\_age\_by\_housing <- credit\_data %>%

group\_by(HOUSING.TYPE) %>%

summarise(Average\_Age = mean(AGE, na.rm = TRUE))

print(average\_age\_by\_housing)

##Intermediate Questions

#Question -1) How does the number of children affect the total household income, and what does this imply for credit risk?

# Ensure necessary libraries are loaded

library(ggplot2)

library(dplyr)

# Calculate average total household income for each number of children

average\_income\_by\_children <- credit\_data %>%

group\_by(NO.OF.CHILDREN) %>%

summarise(Average\_Income = mean(TOTAL.HOUSEHOLD.INCOME, na.rm = TRUE))

# Plot the variation of total household income with the number of children

ggplot(average\_income\_by\_children, aes(x = NO.OF.CHILDREN, y = Average\_Income)) +

geom\_bar(stat = "identity", fill = "pink") +

geom\_text(aes(label = paste0("$", formatC(Average\_Income, format = "f", big.mark = ",", digits = 2))),

vjust = -0.5, size = 3, angle = 90, nudge\_y = 10000) +

labs(title = "Average Total Household Income by Number of Children",

x = "Number of Children",

y = "Average Total Household Income") +

theme\_minimal()

#Question -2) What is the relationship between education level and total household income, and what does this imply for credit risk?

# Ensure necessary libraries are loaded

library(ggplot2)

library(dplyr)

# Calculate average total household income by education level

average\_income\_by\_education <- credit\_data %>%

group\_by(EDUCATION.LEVEL) %>%

summarise(Average\_Income = mean(TOTAL.HOUSEHOLD.INCOME, na.rm = TRUE))

# Plot the average total household income by education level

ggplot(average\_income\_by\_education, aes(x = EDUCATION.LEVEL, y = Average\_Income)) +

geom\_bar(stat = "identity", fill = "red") +

geom\_text(aes(label = paste0("$", formatC(Average\_Income, format = "f", big.mark = ",", digits = 2))),

vjust = -0.5, size = 3) +

labs(title = "Average Total Household Income by Education Level",

x = "Education Level",

y = "Average Total Household Income") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

#Question -3) How does the number of years employed relate to total household income, and what implications does this have for credit risk?

# Ensure necessary libraries are loaded

library(ggplot2)

library(dplyr)

# Plot the relationship between years employed and total household income using a scatter plot

ggplot(credit\_data, aes(x = YEARS.EMPLOYED, y = TOTAL.HOUSEHOLD.INCOME)) +

geom\_point(alpha = 0.5, color = "blue") +

geom\_smooth(method = "lm", se = FALSE, color = "black") +

labs(title = "Relationship between Years Employed and Total Household Income",

x = "Years Employed",

y = "Total Household Income") +

theme\_minimal()

##Advanced Questions

#Question -1) How does the distribution of total household income vary across different income types, and what insights can we draw using density plots?

# Ensure necessary libraries are loaded

library(ggplot2)

library(dplyr)

# Density plot of total household income by income type

ggplot(credit\_data, aes(x = TOTAL.HOUSEHOLD.INCOME, fill = INCOME.TYPE)) +

geom\_density(alpha = 0.5) +

scale\_x\_continuous(labels = scales::dollar\_format()) +

labs(title = "Density Plot of Total Household Income by Income Type",

x = "Total Household Income",

y = "Density",

fill = "Income Type") +

theme\_minimal()

#Question -2) How can we segment the applicants into distinct clusters based on their financial characteristics using K-means clustering?

# Install necessary packages if not already installed

install.packages("ggplot2")

install.packages("dplyr")

install.packages("corrplot")

install.packages("magrittr")

# Load necessary libraries

library(ggplot2)

library(dplyr)

library(corrplot)

library(magrittr)

# Simulate a binary target variable 'DEFAULT' for demonstration purposes

set.seed(123)

credit\_data$DEFAULT <- sample(0:1, nrow(credit\_data), replace = TRUE)

# Prepare the data for correlation analysis

correlation\_data <- credit\_data %>%

dplyr::select(TOTAL.HOUSEHOLD.INCOME, YEARS.EMPLOYED, FAMILY.SIZE) %>%

na.omit()

# Compute the correlation matrix

cor\_matrix <- cor(correlation\_data)

# Plot the correlation heatmap with customization

corrplot(cor\_matrix, method = "color", tl.col = "black", tl.srt = 45,

title = "Correlation Heatmap", addCoef.col = "black", number.cex = 0.7,

cl.cex = 0.7, tl.cex = 0.8, mar = c(0, 0, 2, 0))